Communications of the Association for Information Systems

Volume 8 Article 16

February 2002

Special Issue on the AMCIS 2001 Workshops: Speech Enabled Information Systems: The Next Frontier

Alexander Hars
University of Southern California, alexander.hars@marshall.usc.edu

Communications of the Association for Information Systems

SPECIAL ISSUE ON THE AMCIS 2001 WORKSHOPS: SPEECH ENABLED INFORMATION SYSTEMS: THE NEXT FRONTIER

Alexander Hars
Marshall School of Business
University of Southern California
hars@usc.edu

ABSTRACT

Speech technologies are coming of age. They are applied in an increasing number of mobile, call-center, home, and office settings. They challenge the established Graphical User Interface metaphor and promise to fundamentally alter the way humans conceptualize and interact with computers. This change leads to new requirements for the development of information systems. It also provides new research issues and opportunities for the academic community.

In this article, the main elements of speech technologies will be presented and their applications will be discussed. The article does not focus on technical aspects of speech technologies but is concerned with the business aspects of applying such technologies. The article is based on a workshop at the Americas Conference on Information Systems 2001 in Boston.

KEYWORDS: natural language processing, speech recognition, dictation systems, speech synthesis, interactive voice response, voice XML

I. OVERVIEW

Speech technology has been the subject of intensive research for more than fifty years (Figure 1). Initial enthusiasm was followed by the realization that the challenges are huge. Visions of computers such as HAL [Clarke, 1968] that readily understand human language are still in the distant future. Even optimists don't expect them to be available within the next 10 years. The problem is complex. For example, acoustics alone are not sufficient for speech recognition. Sound profiles of human speech are highly ambiguous and allow many interpretations. Which one is correct can only be determined in context. This in turn requires common-sense knowledge and reasoning – a process at which the human mind excels but which has proven elusive for computers [Lenat, 1995].

Nevertheless, the field has made significant progress [Dahl, 2000; Zadrozny et al., 2000]. In the recent years speech recognition matured sufficiently to become a viable input medium. Speech synthesis has been available for two decades with significant improvements in the last five years. The number of commercial applications multiplied: Speech technology is increasingly used for telephony applications in interactive voice response systems. In the US, several voice portals have been established. In certain professions and industries, computer-based dictation

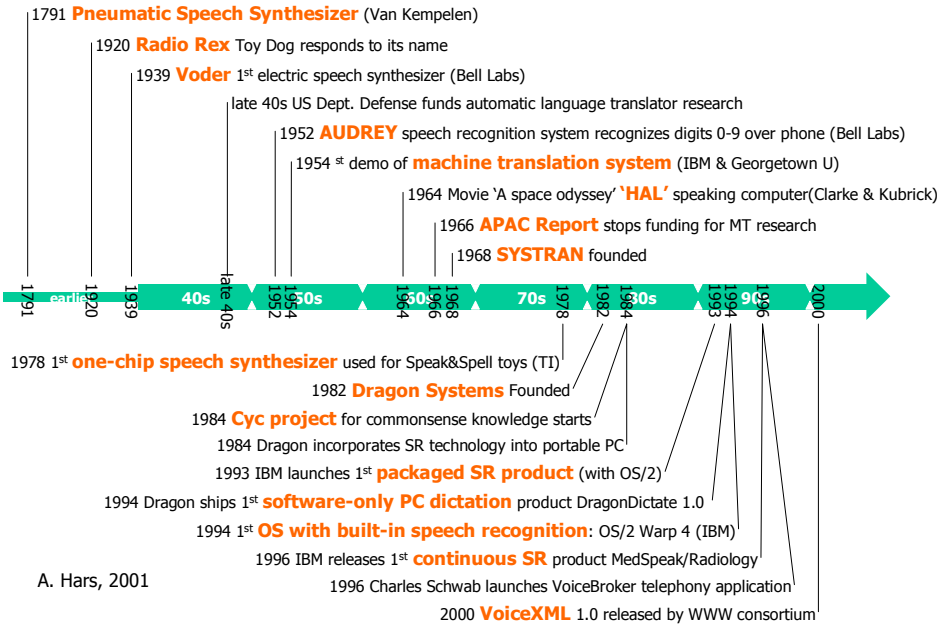

Figure 1. Speech Technology Timeline

systems are used quite regularly. On the Internet, many sites now offer translations services and on several commercial sites, interactive chatterbots or conversation agents offer product advice and engage customers in dialogues.

However, much of this progress has gone unnoticed. Skepticism about the viability of speech technologies hinders further adoption. There are concerns that speech technologies are not yet mature. The disillusionment with AI that began in the 1960's is another hindrance. Users and managers have not yet learned to accept the particular challenges of speech technologies. For example, speech applications tend to be imperfect. They fail part of the time and need ways to recover from errors. This characteristic changes the objectives, structure and design process compared to traditional IT applications. Another problem is the lack of integration of speech technologies into current operating systems platforms. Radical improvements are announced regularly by operating systems vendors and just as regularly postponed to the next release.

This article describes the main speech technologies and their applications. In addition, implications for information systems research (as opposed to computer science research) are presented. The article is organized as follows: The next section presents an overview of current speech technologies. Speech recognition, speech synthesis, and dialogue systems are discussed in Sections III, IV, and V, respectively. Section VI examines the implications for research and practice.

II. ADVANTAGES AND LIMITATIONS OF SPEECH TECHNOLOGIES

Before looking at the details of speech technologies, it is necessary to evaluate the promise of these technologies. What makes speech technologies potentially so attractive and what are their downsides?

COMPONENTS OF SPEECH TECHNOLOGIES

Speech technologies can be broken down into core functional components as shown in Figure 2. Speech recognition – shown on the lower left – translates acoustic utterances into

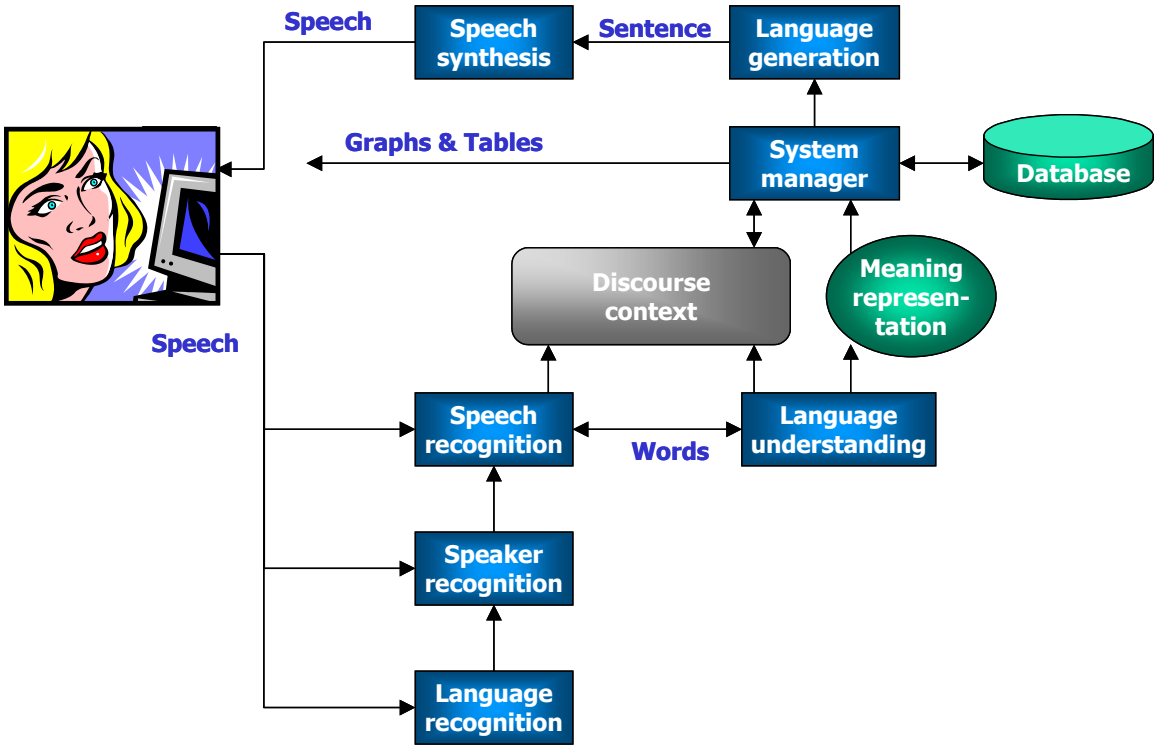

Figure 2. Technologies for Spoken Language Interfaces
[adapted from Zue and Cole, 1996]

symbolic representations. This process translates acoustic signals into internal representations. To do so requires disambiguating the context and can be improved if part of the semantics can be recognized (natural language understanding). Related to speech recognition are two other functions: speaker recognition and language recognition. The former can be used to authenticate who is speaking. The latter is important for multi-lingual applications.

The opposite of speech recognition is speech synthesis. The problem is not just to convert words into sounds, but also to add prosody – intonation and melody that corresponds to the context of the conversation. In most cases, speech synthesis needs to be preceded by language generation. Its task is to translate information that needs to be conveyed into meaningful sentences. Speech technologies also involve a number of additional problems that will not be addressed in this paper. They include language translation, information retrieval [Cowie & Lehnert, 1996] and more specialized technologies such as natural language parsing.

Language understanding and the internal representation of meaning continue to be one of the hardest problems in speech recognition. They hold the key for significant improvements in speech technologies and much computer science research is focusing on this problem. This paper does not concentrate on these issues.

When the advantages of speech technologies are evaluated, they need to be compared to the current interface standards: the graphical user interface (GUI) for output and keyboard and pointing device for input.

ADVANTAGES OF SPEECH INPUT

Ideally, speech recognition is more advantageous than the keyboard because it accelerates data entry for all users with the exception of very prolific typists. In addition, speech is natural and requires no learning while typing requires significant training. Speech recognition also reduces the cognitive load when formulating text. The user can concentrate on the content and need not be distracted by data entry issues. One of the key benefits of dictation is the ability to follow a

train of thought without distraction. Thus speech recognition may significantly increase the productivity of the many knowledge workers who currently type their own reports and articles. In practice, these advantages are reduced by recognition errors, but the accuracy of commercial recognition engines is steadily improving.

Another advantage of speech in- and output are size and energy requirements of the interface. Microphones and speakers require very little power. They are much smaller than even the smallest portable keyboards and display devices. They can be embedded into walls or equipment. As cell phones show, they allow significant shrinking of the size, weight, cost and power consumption of in- and output devices.

Moreover, speech interfaces have the additional advantage that they can be operated without hands and eyes while engaged in other activities. Examples are mobile and hands-free environments.

DISADVANTAGES OF SPEECH INPUT

However, speech input may also have some disadvantages. Some researchers caution that extensive speech input can be tiring. At this point it is not clear whether this limitation is significant. Repetitive typing is fraught with the same problem.

More significantly, speech is not private. Speech input can not be used in meetings or while on the phone when the other party should not become aware of data entry. In addition, high noise levels can reduce accuracy – speech input does not work well in cubicles. But the latter problem may go away as speech recognition improves.

At present, advanced speech recognition is also limited by the need to wear a high-performing microphone. Users are tethered to the computer. Fortunately, this is not likely to be a fundamental issue. Better recognition accuracy and microphone arrays (some of which already have been approved for speech recognition) may eventually eliminate this problem.

ADVANTAGES OF SPEECH OUTPUT

In comparison to speech input, the advantages of speech output are less compelling. Speech output can alert the user and draw attention towards the interface. In contrast to the display, the interface is active and can initiate a dialogue.

Similarly as speech input, speech output also has the advantages of small size, cost and power consumption as well as the ability to perform in mobile and hands-free settings.

DISADVANTAGES OF SPEECH OUTPUT

The bandwidth of information that can be conveyed via the interface is much smaller than what can be conveyed on a display. An average speaker utters about 175-225 words per minute. A reader, in contrast, easily absorbs 350-500 words per minute [Schmandt 1994, p.101]. It is possible to accelerate speech playback electronically to rates of about 300 wpm. For an example see http://www.elantts.com/indemo.htm. But acceleration significantly increases the cognitive load of the listener.

In addition, speech output is linear and sequential. Whereas displays present many pieces of information in parallel, a speech interface can only convey one item at a time. Some researchers attempted to use the excellent spatial capabilities of the ear for interfaces involving three dimensional soundscapes. While these approaches can provide additional cues, it cannot eliminate the fundamental limitation of sequential spoken language.

Associated with the spatial representation of information on displays is the existence of pointing devices that allow a user to refer to elements that are part of the context. Pointing devices greatly simplify the interaction with graphical user interfaces. While it is true that speech interfaces currently lack pointing devices, it is not adequate – as many authors claim – to argue that speech interfaces inherently lack pointing devices. While the mouse is probably not useful, a stick that points e.g. to the right for forward, fast forward, and to the left (rewind, fast rewind) would clearly be useful. It is very much conceivable, that additional pointing features could be added to switch between application contexts (e.g. when performing several different interactions in parallel such as switching between dictating several letters, listening to voicemail, checking the status of some external information, etc). Finding the best approaches and metaphors are challenges for interface design, rather than fundamental limitations of the speech interface.

Table 1 summarizes the advantages and disadvantages of speech input and output.

Table 1. Advantages and Disadvantages of Speech Input and Output

	Advantage	Disadvantage
Speech Input	Accelerates data entry	Can be tiring
	Size and energy required	Not private
	Operated without hands and eyes	Wearing microphones
Speech Output	Small size	Bandwidth required
	Cost	Linear
	Power consumption	Sequential
	Mobile and hands-free settings	Lack of pointing devices

EVALUATION

It follows that speech interfaces can be useful in many situations. In mobile environments, in processes where interaction occurs intermittently and where the system needs to draw the attention of the user, whenever large amounts of texts need to be entered, then speech interfaces have clear advantages. On the other hand while speech input may eventually become the by far predominant input technology, speech output will certainly not replace the display. Speech output has significant advantages, in particular the size requirements, location independence, and the ability to engage in other activities while absorbing information. Thus it will grow in importance, but it will neither displace the screen nor relegate the screen to niche status.

III. SPEECH RECOGNITION

Much progress has been made in speech recognition. Five years ago, commercial speech recognition was viable only for niche markets with highly specialized vocabularies or very small vocabularies as used in interactive voice response systems. Before dictation systems could be used, at least 30 minutes had to be spent in adapting the speech recognition engine to the speaker. Speakers needed to make a short pause between every word. Today, the two leading commercial dictation systems (Lernaut & Hauspie's Naturally Speaking (originally from Dragon Software) and IBM's ViaVoice) routinely process continuous speech. They require less than 5 minutes for the first adaptation and their accuracy is greatly improved. This progress is not only a result of better speech recognition algorithms, it is also due to the average desktop's improvements in computing performance. Speech recognition is very computing-intensive.

Currently five criteria need to be examined to evaluate and compare speech recognition systems:

- **Vocabulary size:** Speech recognition engines differ greatly in the size of the vocabulary supported. Custom speech applications used in telephony settings or for hands-free data entry have very limited vocabularies. Many speech applications only need to recognize digits, some additional number-related terms, yes and no and a few commands. A small vocabulary greatly reduces performance requirements and allows recognition without training. It allows embedding limited vocabulary engines into hand-held devices. Dictation systems, on the other hand, require much larger vocabularies. Current systems typically recognize more than 200,000 words and word variations. Progress is rapid and it will not take long until dictation systems support the complete vocabulary of a language.
- **Resource requirements:** The resource requirements are another major issue. They are very much tied to the size of the vocabulary. Dictation systems require high performing PCs with a large memory. They are not feasible on limited hardware such as palmtops and wireless devices. Limited vocabulary recognition however is feasible in such environments.
- **Speaker dependence:** Speaker dependence can be measured by the amount of training that a recognition engine requires to perform well for a targeted speaker. Current general-

purpose dictation systems require very little training time. Training time may also be required in noisy environments.

- **Accuracy**: Accuracy applies to the quality of speech recognition systems under ideal conditions. Current speech recognition systems report accuracies of about 95 percent.
- **Robustness:** Robustness reflects the accuracy of a recognition engine under less then perfect conditions. Many factors reduce the accuracy in real-world settings: A good sound card and a noise-canceling microphone (correctly worn close to the mouth) are essential. Ambient noise and heavy accents further degrade robustness.

At the current level of accuracy speech recognition software is becoming a viable alternative to the keyboard. Even with an error rate of about 10 percent, many users will be able to increase the speed at which sentences are entered. Overall, speech recognition systems have matured greatly. As processor performance increases dictation systems will rapidly become a serious alternative for data entry. The one disadvantage that will persist in the medium term is the necessity to carry a headset.

APPLICATIONS OF SPEECH RECOGNITION

For the last several years, speech recognition has been applied in real-world, commercial settings. In the following, three typical application areas of speech recognition are discussed:

- medical dictation,
- voice authentication and
- speech recognition in a warehousing environment.

Speech recognition also plays an important role in telephony applications (interactive voice response) that will be discussed in Section V.

Medical Dictation

Medicine was one of the first areas where speech recognition took hold. Physicians use a highly specialized vocabulary with many distinctive terms consisting of many syllables. They are required to document their observations and conclusions in short semi-structured notes. Traditionally many physicians dictated their reports into a recorder. A transcription department typed the document and returned it to the physician. He or she then reviewed and edited the document before signing off on it. It could be a lengthy process.

As a consequence, some hospitals including the radiology department at Duke University implemented speech solutions [Dictaphone, 2001]. The department's approximately 100 staff physicians and residents produce about 340,000 reports per year. Before implementing speech recognition, reports were turned around in approximately 48 hours. Duke Radiology then rolled out technology from Dragon Software (since acquired by Lernaut & Hauspie) in which physicians dictate their reports into a digital recorder. The recorder can be plugged into the network where it automatically uploads its contents to a speech server. The recorders can also be plugged into medical PCs. The system automatically transcribes the text and displays the result on-screen. The radiologists can make any correction and then sign the document. Speech recognition reduced the cycle time for reports to about 4 hours.

Nevertheless, Duke Radiology did not force physicians to use the system. The traditional process is still supported. Although physicians can still have their dictation transcribed by professional transcriptionists, only five percent of the radiologists use this alternative. Ninety-five percent now edit the digital transcription themselves. Beyond reducing cycle time, the speech recognition system also reduced the error rate. The different workflows that are typically encountered in report transcription are shown in Figure 3. As will be seen from other applications discussed in this paper, it is typical for speech applications that they do not completely eliminate the traditional process.

Continuous dictation systems are also used in other industries, especially in the legal profession, in the insurance industry and wherever forms need to be processed.

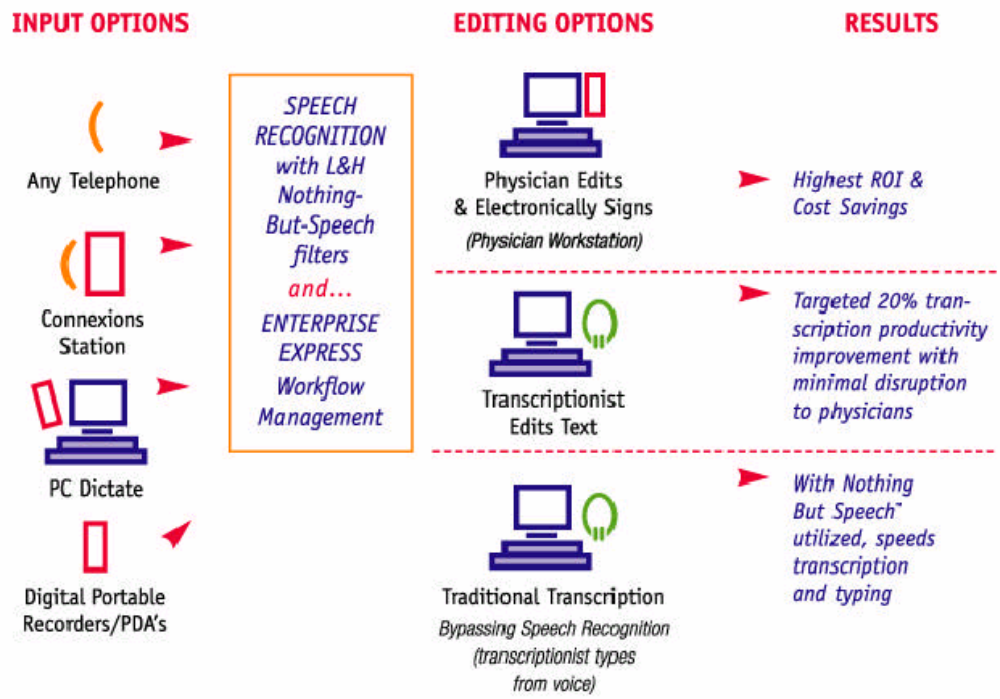

Figure 3. Medical dictation workflow [Lernaut & Hauspie, 2001]

Inventory Management

Speech recognition is also well suited for mobile and hands-off environments. A classical example is warehouse applications where workers need to move items or check the status of items. In a typical scenario [Weinschenk & Barker, 2000, pp.167-170] a speech application instructs workers to fill customer orders. The workers are equipped with radio-controlled headsets with microphones. When a new order needs to be filled, they receive an audio instruction such as

(System): "go to aisle six, shelf five, bin two, pick four"
Once they have completed this assignment, the picker confirms
(Worker): "picked 4".

Such a system increases productivity by eliminating the need to carry a notepad while walking through the aisles and by eliminating the time to stow the notepad when retrieving items with both hands. In addition, the system can react to problems (e.g. when no more items are available). The vocabulary in such systems tends to be very small but recognition accuracy can be impaired because of high noise levels typically present in warehouses.

Voice Authentication

Another application of speech recognition is voice authentication. Every voice is unique and thus can be used for identification. In theory uniqueness has great advantages: it obviates the need for passwords, thereby eliminating the danger of losing a password. In addition, passwords can not be passed on from authorized to unauthorized users. However, a study conducted by the Center for Communication Interface Research at the University of Edinburgh [2000] for Nuance, a provider of such technology, found significant limitations. In the study a banking application was simulated. One thousand participants were asked to identify themselves using 19 spoken digits. They contained a member number, an account number and a two-digit pin. Even with this long sequence of digits, the verification error rate was 1.2 percent. This rate is not yet acceptable for commercial applications. In addition, the error rate for identical twins was much higher at 50 percent. Thus while verification holds some promises, it is far away from being able to replace passwords.

IV. SPEECH SYNTHESIS

The second major component of speech technologies is speech synthesis. First attempts to produce machines that speak go back to the 18th century. Van Kempelen developed a pneumatic speech synthesizer in 1791 [Manell, 1998]. Today, three major approaches for speech synthesis are being used.

Concatenative Synthesis. Many interactive voice response systems use recorded text that they assemble on demand. In this approach, which is called concatenative synthesis, a trained reader records utterances (words or sentences). The digital utterances are then combined by the system as necessary. This approach leads to natural sound and minimizes performance requirements. However large amounts of memory are needed for storage. The approach is only suitable for limited vocabularies. We are all familiar with this approach from phone mail systems.

Diphone Synthesis. A more elaborate version of this approach is diphone synthesis. Words are broken down into the 44 to 48 phonemes which are relevant for English. Phonemes reflect the way in which consonants and vowels can be spoken. Next all possible combinations of these phonemes are spoken and recorded. These combinations are called diphones. The about 1500 to 1800 diphones represent every combination of two letters. When sentences are processed, they are converted into diphone sequences. Next, intonation is added. The output of diphone synthesis can sound monotonic. But much recent work has focused on improving intonation. Most current speech generators, including the currently most advanced systems from Lernaut & Hauspie and from ATT (Table 2)) rely on diphone synthesis.

Formant Synthesis. A different approach for speech synthesis is formant synthesis. It relies on a mathematical model of the human speech apparatus, which contains aural chamber, trachea, larynx etc. Text is then converted into tongue and lip movements. This approach produces very natural sound. Another advantage is that it allows voice formatting. For example, the aural chamber can be enlarged to produce deeper sound for emphasis.

Most companies that offer speech synthesis tools have websites that demonstrate their capabilities. Table 2 contains URLs of the most advanced speech synthesis products:

Table 2. Online Speech Synthesis Demonstrations of Leading Vendors

Company / Product	URL
Lernaut & Hauspie Realspeak	http://www.lhsl.com/realspeak/demo.cfm
ATT NaturalVoices	http://www.naturalvoices.att.com/demos/
Lucent Bell Labs Articulator	http://www.bell-labs.com/project/tts/voices.html
Elan SpeechCube	http://www.elantts.com/indemo.htm

V. DIALOGUE SYSTEMS

Speech recognition and synthesis technologies together form the basis for systems supporting natural language dialogues. In this section, technologies that support spoken dialogue will be presented, in particular systems for interactive voice response. In the future, dialogue applications will also be available on the desktop. They are also emerging for wireless devices.

VOICE XML

Recently VoiceXML emerged as a new standard for interactive voice response (IVR) systems. This standard is supported by most vendors of IVR systems and by the World Wide Web Consortium. The goal of voice XML is to provide a common language for content providers, tool providers, and platform providers. It shields authors of voice applications from application and hardware details. Voice XML supports standard telephony interactions such as voice menus and prompts. It provides excellent support for basic dialog features and allows easy extension through software and scripts. Voice XML currently is available in version 1.0.

An example of a voice XML form is shown in Figure 4. It specifies a voice menu that offers several choices. The “prompt” command contains text that the system needs to

synthesize. The "enumerate" option ensures that the individual choices are spoken by the computer. Each choice contains a link to the next voice template for further processing. Voice XML also contains special commands for handling errors. The “no match” element is triggered when the system does not recognize the user's answer. When the system does not register any response, it follows up with the content of the “no input” element.

```
<?xml version="1.0"?>
<vxml version="1.0">
    <menu>
    <prompt>Would you like <enumerate/></prompt>
      <choice next="http://..coffee.vxml">coffee</choice>
      <choice next="http://..tea.vxml">tea</choice>
      <choice next="http://..milk.vxml">milk</choice>
      <choice next="http://..nothing.vxml">nothing</choice>
      <nomatch>I didn't understand what you said.</nomatch>
      <noinput>You must say something.</noinput>
    </menu>
</vxml>
```

Figure 4. Voice XML Example [Rehor et al, 2000]

Voice XML supports several ways of generating output. One approach is speech synthesis. VoiceXML also supports prerecorded audio files. At present three input modes are supported: the system can

- use speech recognition based on an inline or external grammar,
- record input as audio files, and
- accept standard tone signals from the telephone.

In the future VoiceXML systems are also expected to allow the transcription of text.

Moreover VoiceXML supports several different dialogue styles including:

- traditional prompt menus.
- alternatively directed dialogues where the user is prompted for voice response.
- (most advanced) mixed initiative dialogues.

Mixed initiative dialogues allow the user to change the dialogue context and to jump to a different type of dialogue. To do so, the system must listen for words that indicate a different topic and then switch to the topic. Voice portals provide examples of this technique. For example, a user can logon and retrieve information about national news. While the system plays this information, it listens to interruptions by the user. For example, the user may say "go to sports". The system recognizes that this command refers to a different VoiceXML template. It stops playing the current information and proceeds to the sports section. This feature has great advantages because it makes voice interaction much more natural.

Voice XML also provides telephony features. Voice XML systems can initiate phone calls. For example, a financial information system can alert its customers to changes in the stock market (Figure 5). When an opportunity arises it can dial of the phone number of a customer. When the customer picks up, it informs the customer about the market situation and proposes default actions. Similarly voice XML systems can put callers on hold or they can transfer a call to a customer representative. Table 3 provides URLs to online demonstrations of voice XML systems.

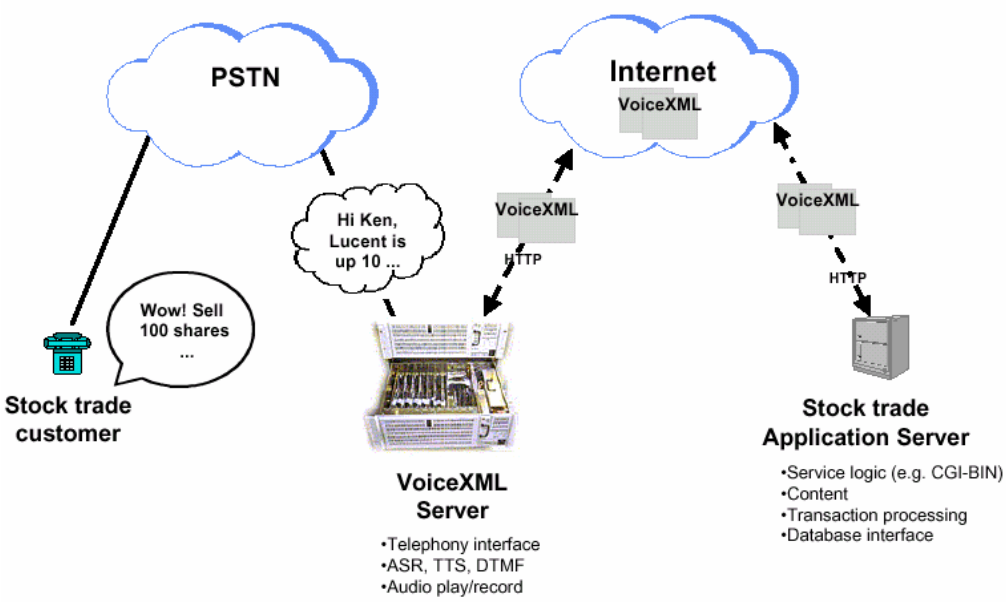

Figure 5. Dialogue Initiated by a VoiceXML System [Rehor, et. al, 2000]

Table 3. Selected Demonstrations of Interactive Voice Response Systems

Vendor	Application type and URL
General Magic	Credit Security – Reporting lost or stolen credit cards http://www.generalmagic.com/aboutvoice/credit_security_demo.ram
Nuance	Banking Demo – Dial 650-847-7438 http://www.nuance.com/demos/demo-banking.html
Speechworks	United Flight Information http://www.speechworks.com/demos/travel.cfm

Nuance [2001] also provides an indication of the costs associated with an interactive voice response system for a medium-sized call center (Table 4). The system provides 72 simultaneous connections which – taking into account demand fluctuations – is equivalent to approximately 50 call center agents. The cost structure shows that significant effort is required for developing, integrating and testing the customized call center application. In addition, maintenance incurs almost fifty percent of the total costs.

Table 4. Cost Structure of a Medium-sized IVR Application [Nuance, 2001]

Item	Cost
72 Port IVR Hardware	$164,000
Speech Software	180,000
Application Development	95,000
Implementation	55,000
Annual Maintenance	114,000
Total (4 years)	$950,000

Speech Portals

With the growing interest in the wireless web, several voice portals were established. Voice portals provide information services via the phone. A list of voice portals is shown in Table

5. Most of these portals were established to showcase portal technologies and can be accessed free of charge. The portals provide information about news, stocks, sports and the lottery. They provide access to email. Some also provide restaurant recommendations and driving directions. TellMe offers a location-based service that connects to a local taxi company for arranging transportation. In most portals, users can set their own preferences – either through a voice dialogue or through a web based interface.

Table 5. Voice Portals

Portal	Phone Access	URL
Tell-Me	800-555-Tell	www.tellme.com
BeVocal	800-4BVOCAL	www.bevocal.com
HeyAnita	800-44-ANITA	www.heyanita.com
AudioPoint	888-38-AUDIO	www.myaudiopoint.com
TelSurf	818-87-41280	www.888telsurf.com

VI. RESEARCH CHALLENGES

Speech technologies have matured sufficiently to be used in many commercial settings. It is time to examine speech technologies not just from a computer science perspective but also from the perspective of information systems research. This section discusses several challenges for application oriented and information systems research.

EVALUATION METRICS

Speech technologies are imperfect technologies. They have significant error rates and are often not reliable enough to replace traditional processes fully. These shortcomings increase process complexity and leads to significant risks and tradeoffs. Therefore criteria and metrics are needed that can be used to evaluate speech solutions.

USER INTERFACE DESIGN

Speech technologies provide major challenges for user interface design. When only speech input and output is available, it is necessary to rethink the interaction between user and system. While it is possible to incorporate many user interface elements such as selection lists, radio buttons, application contexts (e.g. windows), (audio) icons etc. that correspond to elements from the established GUI metaphor, traditional applications can not be ported easily to speech interfaces [Raman, 1997]. Innovations are needed which account for the critical role of context that is often implicitly established in a natural language dialogue.

While GUI-based applications typically leave the initiative with the user, speech-based applications need to support mixed-initiative dialogues. For example, applications often do not wait to be invoked by the user; they request the user's attention. Furthermore, speech applications often make inferences from the users' responses that determine what further options are presented to the user and what default actions the system may take. Incremental learning and the adaptation of the user interface to implicit and explicit user characteristics become much more important [Browne, Totterdell & Norman, 1990].

A second challenge in user interface design lies in merging speech technologies with traditional GUI applications. It is not clear, for example, what kinds of messages a system should speak when the user is already glued to a screen. Many interesting questions arise when dialogues can be multi-modal.

SPEECH TECHNOLOGY LIFE-CYCLE AND DEVELOPMENT APPROACHES

It is necessary to examine how the life-cycle of speech technologies differs from traditional information systems life-cycle. Clearly, speech technologies require iterative development with heavy prototyping. Speech applications require constant monitoring and adaptation, and they have a very active maintenance phase.

PROCESS IMPLICATIONS

Speech technologies provide new opportunities for business process redesign. They extend the reach of information systems to workers in remote locations, in moving and hands-free environments. In contrast to most traditional information systems applications, they are able to establish communication to workers and customers autonomously by initiating a phone call. They can broadcast messages to larger groups or sample opinions in a short time. Thus speech technologies provide ample capabilities for process innovation.

BUSINESS MODELS FOR SPEECH TECHNOLOGIES

Speech technologies may give rise to new business models. Several enterprises already established themselves as voice portals – although their economic viability is not yet clear. Speech technologies may also lead to restructuring and increased outsourcing of call center organizations. Speech technologies may also give rise to new types of information and infomediary services.

SOCIAL IMPLICATIONS

Important implications also result for society. How dependent will we be on writing in the future? Will our children need to learn how to write? Will the emphasis of education change if writing no longer is a key skill? Finally, how will work structures change as call centers become self-service centers and as computers take a more active part in customer interaction? Speech technologies may change the way we think about computers. If we interact with them in a natural way, will we continue to think about them as mindless machines? Speech interaction, much more than any other software technology, will challenge our view of intelligence and with it our view of ourselves.

VII. CONCLUSION

Almost unnoticed, speech technologies are moving into the mainstream, and an increasing number of companies are adopting these technologies. Although speech interfaces offer many advantages over traditional interfaces, they provide many challenges; in particular, the interaction between human and computer has to be redefined. In addition, much research is necessary to develop new interface paradigms that are able to take advantage of context and user models. Further, business processes will be greatly changed, as workers are untethered from their desktops and laptops.

Ultimately, speech interfaces will redefine not only our interaction with the computer but also our concept of computing. Will the idealization of computing move from the PC that dominated the eighties and nineties after replacing the mainframe (fifties to seventies) to the ubiquitous voice? Speech-based information systems are poised to become the next frontier in information systems.

Editor's Note: This article is based on the author's workshop presented at AMCIS 2001. It was received on September 14, 2001 and was accepted on October 18, 2001. The article was with the author for approximately 5 weeks for three revisions. The article was published on February 27, 2002 together with the other articles in the special issue on the AMCIS 2001 Workshops.

REFERENCES

EDITOR'S NOTE: The following reference list contains the address of World Wide Web pages. Readers who have the ability to access the Web directly from their computer or are reading the paper on the Web, can gain direct access to these references. Readers are warned, however, that

1. these links existed as of the date of publication but are not guaranteed to be working thereafter.
2. the contents of Web pages may change over time. Where version information is provided in the References, different versions may not contain the information or the conclusions referenced.
3. the authors of the Web pages, not CAIS, are responsible for the accuracy of their content.
4. the author of this article, not CAIS, is responsible for the accuracy of the URL and version information.

Browne, D., P. Totterdell, and M. Norman (1990) *Adaptive user interfaces.* London:Academic Press.

Centre For Communication Interface Research, Univ. of Edinburgh (2000) *Large scale evaluation of automatic speaker verification technology*. Technology Report. http://www.nuance.com/pdf/ccir_execsum.pdf (7/20/2001).

Clarke, A.C. (1968) *2001: A Space Odyssey.* London: Hutchinson.

Cowie, J. and W. Lehnert (1996) "Information extraction". *CACM*, (39)1, pp. 80-91.

Dahl, D.A., L.M. Norton, and K.W. Scholz (2000) "Commercialization of NLP technology". *CACM*, (43)11.

Dictaphone (2001) *Duke University Medical Center Turns To Dictaphone PowerScribe*. Online Case Study. http://www.dictaphone.com/healthcare/ case_studies/dukcs.asp (7/18/2001).

Hars, A. (2000) "Web-based knowledge infrastructures for the sciences: an adaptive document" CAIS (4)1

Lenat, D.B. (1995) "CYC: A large scale investment in knowledge infrastructure". *CACM* (38)11, pp. 32-38.

Lernaut & Hauspie (2001) "Dictaphone Enterprise Express". Brochure. ftp://206.26.152.6/pdf/pdf_mkt/eexbro.pdf (7/15/2001).

Manell, R.H. (1998): *Brief historical introduction to speech synthesis.* http://www.ling.mq.edu.au/~rmannell/slp807/history_synthesis/ (8/20/2001).

Raman, T.V. (1997) *Auditory user interfaces*. Boston: Kluwer.

Rehor, K.G., P.J. Danielsen, and C. Tuckey (2000) Building speech recognition telephony applications with Voice Extensible Markup Language. Presentation Slides. http://www.voicexml.org/avios2000_voicexml.pdf (8/20/2001).

Schmand, C. (1994) *Voice communication with computers*. New York: Van Nostrand Reinhold.

Weinschenk, S. and D.T. Barker (2000) *Designing effective speech interfaces*. New York: Wiley .

Zadrozny, W. et al. (2000) "Enabling technologies: natural language dialogue for personalized interaction". *CACM* (43)8, pp. 116-120.

Zue, V. and R. Cole (1996) "Spoken language input: Overview". In Cole, R. et al. (eds.): *Survey of the state of the art in human language technology*. http://cslu.cse.ogi.edu/HLTsurvey/ch1node3.html#SECTION11 (1/15/2001).

ABOUT THE AUTHOR

Alexander Hars is Assistant Professor of Information Systems at the Marshall School of Business, University of Southern California. He specializes in knowledge management, enterprise modeling, and applications of speech technologies. His current research interests are knowledge-based analysis of information systems, business reference models and conversation systems for requirements engineering. He is the founder of an innovative infrastructure for information systems research (cybrarium.usc.edu) that experiments with new approaches for electronic networks of scientific knowledge [Hars 2000]

Communications of the Association for Information Systems

ISSN: 1529-3181